On Multivariate Analysis*

BU-50-M R. G. D. Steel June, 1952

Professor Snedecor, in a chapter on covariance in his text, deals with the yield by weight and number of beets in a randomized block experiment. Here he analyzes each variable separately and then goes on to a test of yield (weight) treatment means when adjusted for stand. These tests, he relates to one another pointing out certain features of such data and how they may be taken into consideration in summing up the results and in drawing conclusions. One of the points he raises about covariance in general is whether its use is required at all or whether a single variable may not give all the relevant information. "It is for the investigator to decide what question he is proposing". The possession of information on several variables is not a sufficient reason for the use of covariance. The important question when information on more than one variable is available, providing one variable does not supply all the necessary information, is how to summarize the data to the best advantage, making full use of the available information contained in the interrelations among the variables.

Professor Hotelling has just discussed techniques which use information from several variables, techniques which are not covariance, techniques which give us overall tests of significance

- - - - - - - - - - - - - - - - -

* A talk presented at the Biostatics Conference at Ames, Iowa, June 1952, following that of Professor Harold Hotelling.

rather than tests on a single variable, adjusted or unadjusted, and techniques which combine data from several variables and allow us to construct confidence intervals in some instances. (We are also acquainted with path coefficients, multiple and partial regression, and analysis of variance and covariance.)

Tests of hypotheses are essential for inference. In order to assign a measure of credibility to our inference, it is necessary to do more than simply describe a test. With multivariate data, the problem is more than some combination of univariate tests. For example, with two variables we might prepare to claim a significant difference between the means of two populations on the basis of our sample if a test of treatment means for either one of the variables was declared significant at the 5% point by use of Student's "t" as a test criterion. If the two variables are completely dependent, then our test is truly at the 5% point; however, if the variables are completely independent, then the probability of claiming a difference when it does not exist is .0975, almost 10%, and for independence in general, the probability of failing at least one test is $1 - (.95)^n$. On the other hand, if we agree to reject only if tests on means for both variables show significance, then with independence our level is $.05^2 = .0025$ and it becomes difficult to detect all but the grossest differences, and for independence in general, the probability of failing all tests is $(.05)^n$.

Though independence of sub-tests seems to imply a difficulty, yet such tests furnish more information when considered jointly. Professor Hotelling has been talking, among other things, about functions that give us joint measures of the departure of all observations from their

expected values and by means of which we can attach measures of tenability to our hypotheses. If such a comprehensive test indicates no significance, we are through for the time being at least; if it shows significance, then we may wish to consider a variable by variable analysis to locate the cause and help the experimenter. I do not mean to imply that this will give a complete answer. A complex experiment must be expected to have a complex answer and may have one - as many of you who deal with factorial experiments know. Alternatively, one may wish to use a discriminant function, either statistical or economic. Interpretation will depend to some extent on the nature of the treatments. In a univariate case, we may consider main effects and interactions, other methods of using individual d.f. such as testing for linearity, or a combination of techniques such as Tukey has proposed.

To illustrate two of the simpler tests available, I have some data from a pilot plant fermentation study with yeast. The experiment consisted of a single 6x6 Latin square in which at least 13 variables were measured. Of these variables, it had already been decided that some did not supply the information that they were meant to and perhaps not any information, significant differences within other variables seemed to have no importance as a satisfactory range was all that was required. Finally, we ended with 4 variables: i) lbs. at the end of the first period, ii) lbs. at the end of the second period, iii) fresh activity and iv) exposed activity. (A partial summary of the data is available on a separate sheet. This is in the form of sums of squares and cross products.)

Now, clearly there can't be too much profit in increasing the yields if they are negatively correlated with the activities. Also,

information about lack of correlation or the existence of positive correlation is valuable. Hence, let us test the independence of these two sets of variables. This is done by evaluating the determinant of the variance-covariance matrix of residuals (or sums of squares or correlations) for all four variables and dividing it by the product of the determinants of each of the two sets. (I used sums of squares to avoid an operation and the possible introduction of rounding errors.) The square root of this quantity has the beta-distribution (related to F) with parameters p and q equal to (error d.f. + 1 - no. of variables) and no. of variables in the larger set, viz. 17 and 2. The probability of a smaller and more discrepant value than .773, the value obtained, lies between .06 and .07. The evidence would seem to be in favor of obtaining more information. The evaluation of a 4x4 determinant is not difficult and those of you who run regressions with 4 independent variables do just that in the process of testing the multiple correlation coefficient. It is interesting to note the apparent lack of correlation within each group. It would seem to indicate that the example is not trivial as it might be if the correlations were pronounced. Of the correlations between the variables of different groups, one is significant at the 5% point. This test makes use of the information of the within and between groups correlations.

For comparison with the usual F-tables, use

$$F = \frac{m}{n} \frac{(1 - w)}{w}$$

where w is the square root described above and $m = 2p$ and $n = 2q$ are to be considered as d.f. for lesser and greater mean squares respectively for the purpose of entering the table. Here $F = 2.5$, $m = 34$ and $n = 4$.

Had there been a highly significant dependence between groups, then one would certainly have considered the possibility of using canonical correlations for prediction purposes. It is doubtful if that would have been desirable here unless the activity tests required more space or time than was usually available, but this is a possibility with the exposed activity variable. The lack of a significant correlation also indicates that all variables would be desirable in a discriminant function. The use of control charts and a T statistic or statistics combining the evidence of at least two of the variables would seem profitable since standard yeasts are run virtually every time a new yeast is on trial. The T statistics aid in making a decision when the individual results seem in conflict. An exact distribution is presently available for two variables, approximations for larger values are available. The exact distribution makes use of past data used in preparing the control charts.

Also, I wish to show tests when two variables are measured. (When there are more than two treatments as in this case, the tests will probably be approximate and significant differences possibly difficult to interpret if more than two variables are used.) The criterion for testing differences among the six pairs of treatment means consists of the ratio of the 2x2 determinant of the sums of squares and cross products for error to the same for error + treatments. The square root of this quantity has the beta-distribution with parameters p and q equal to (error d.f. - 1) and treatment d.f. respectively. The probability of a smaller and more discrepant value than that obtained, viz. .249, is very small ($< 10^{-7}$). This is not surprising as F values for the individual analyses are well beyond the 99% point and the variables are negligibly correlated in the sample.

In the case of the activity variables the probability is of the order of 2 or 3 in 1,000, the square root being .528. Individual analyses show significance at 5% and 1% respectively and again the variables are negligibly correlated.

For use with F tables, calculate

$$F = \frac{m}{n} \frac{(1 - w)}{w}$$

where w is the above mentioned square root and $m = 2p$ and $n = 2q$ for p and q previously defined.

In summary, many tests of multivariate hypotheses and techniques utilizing multivariate data exist. Such tests make use of the information contained in correlations among the variables. Approximate and some exact and familiar distributions are available to determine significance levels and construct confidence limits. (Two examples of tests for which exact distributions are available have been given here.) The tests are of excess joint deviations from expected values. The treatment of data yielding a significant value of the test criterion must depend on the questions asked by the experimenter. (Such treatment has not been attempted above.) Some suggestions are offered.

SELECTED REFERENCES

Hotelling, H. The generalization of Student's ratio, Annals of Math. Stat., Vol. 2, 1931, pp. 359-378.

Hotelling, H. Relations between two sets of variates, Biometrika, Vol. 28, 1936, pp. 321-377.

Hotelling, H. Multivariate quality control, Techniques of Statistical Analyses, McGraw-Hill, 1947, 113-184.

Hotelling, H. A generalized T test and measure of multivariate dispersion, Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, Univ. of Calif. Press, 1951, 23-41.

Hotelling, H. Multivariate Analysis. Mimeograph for Biostatics Conference, Ames, Iowa, 1952.

Snedecor, G. W. Statistical Methods, 4th Ed., Iowa State College Press, 1946, Sections 12.7 and 12.8.

Wilks, S. S. Test criteria for statistical hypotheses involving several variables, Jour. Amer. Stat. Ass'n, Vol. 30, 1935, pp. 549-560.

Wilks, S. S. Mathematical Statistics, Princeton University Press, 1946, Chap. XI.

Wilks, S. S. On the independence of k sets of normally distributed statistical variables, Econometrica, Vol. 3, 1945, pp. 309-326.

Wilks, S. S. The analysis of variance for two or more variables, Report of the Third Annual Research Conference on Economics and Statistics, June 28 to July 23, 1937, Cowles Commission for Research in Economics, Colorado Springs, U.S.A., pp. 82-85.

TABLES

Tables of the Incomplete Beta-Function. Edited by Karl Pearson. Published by the "Biometrika" Office, University College, London.

ERROR SUMS OF SQUARES

and

CROSS PRODUCTS

(Numbers in brackets are correlations)

(d.f. = 20)

	x_1	x_2	x_3	x_4
x_1	6.13	1.013 (.037)	-8.8 (-.372)	4.0 (.176)
x_2	1.013 (.037)	120.71	-53.3 (-.508)	-3.53 (-.035)
x_3	-8.8 (-.372)	-53.3 (-.508)	91.2	-2.9 (-.033)
x_4	4.0 (.176)	-3.53 (-.035)	-2.9 (-.033)	83.88

ERROR + TREATMENT

SUMS OF SQUARES AND CROSS PRODUCTS

(d.f. = 20 + 5)

	x_1	x_2	x_3	x_4
x_1	58.12	100.733	52.6	49.4
x_2	100.733	379.33	89.5	142.07
x_3	52.6	89.5	178.1	76.9
x_4	49.4	142.07	76.9	187.02

x_1 = yield at end of first stage

x_2 = yield at end of second stage

x_3 = fresh activity

x_4 = exposed activity

Data courtesy of Red Star Yeast and Products, Milwaukee, Wisconsin